Analyse statistique de signaux acoustiques environnementaux pour la détection d'événements sonores

Erwann BETTON-PLOYON[1, 2] Abbes KACEM[1] Jérôme MARS[2] Nadine MARTIN[3]

[1]ACOUSTB, 24 Rue Joseph Fourier, 38400 Saint-Martin-d'Hères, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France

[3]ASTRIIS, 710, avenue de la Motte-Servolex, 73000 Chambéry, France

Résumé – La détection d'événements sonores est une étape cruciale dans la caractérisation d'une source de bruit, en vue de calculer sa contribution au sein d'un environnement complexe. Cette détection peut être réalisée avec des techniques d'apprentissage supervisé dont les performances dépendent fortement de la taille et de la qualité de la base d'entraînement. Le travail présenté dans cet article concerne la proposition d'une méthode de détection qui s'appuie sur l'évolution temporelle de quatre critères statistiques. Ces critères sont calculés à partir des niveaux sonores en bandes de tiers d'octave. Ils sont ensuite appliqués pour détecter des ruptures dans le signal, et en déduire la présence d'un événement. La validité des ruptures détectées est étudiée sur une base de signaux sonores. Les résultats illustrent l'intérêt de l'approche proposée. Basés sur l'évolution des moyennes par bande, les critères T^2 et distance euclidienne semblent adaptés à la détection d'évènements sonores large bande de type ferroviaire. Pour atteindre des performances équivalentes, les critères BIC et Kullback-Leibler nécessitent une restriction de la bande fréquentielle.

Abstract – Sound event detection is a critical stage to characterize an acoustic source, in order to calculate its contribution within a complex environment. Supervised learning methods can perform this task, but they highly depend on the dataset accuracy. In this paper, the proposed approach uses the temporal evolution of four statistical criteria to perform sound event detection. These criteria are computed on third-octave bands sound levels of the signal. Criteria detect change points in the sound levels evolutions, indicating a sound event occurrence. Results on our dataset show that this method can detect change points within complex environments. Based on the evolution of mean third-octave bands, T^2 and euclidean distance are more likely to detect broad-band events, like railway traffic. To get equivalent results, BIC and Kullback-Leibler distance need some information on characteristic frequencies.

1 Introduction

Selon le rapport de l'Organisation mondiale de la santé publié en 2018, le bruit constitue le second facteur environnemental, derrière la pollution atmosphérique, provoquant le plus de dommages sanitaires en Europe [1]. L'exposition de longue durée à des niveaux de bruit élevés peut conduire à des dommages irréversibles (surdité, troubles cardiovasculaires, etc.). Plusieurs sources sonores sont ainsi soumises à des textes réglementaires pour lutter contre les nuisances associées.

Pour cette raison, il est important de déterminer, dans un environnement sonore complexe, la contribution des sources acoustiques actives. Chaque source étant potentiellement soumise à une réglementation précise, il faut être capable de détecter les périodes durant lesquelles une source émet.

Pour réaliser cette tâche, il est courant d'utiliser des réseaux de neurones, par apprentissage sur des données labellisées. Ensuite, ces réseaux sont capables de renseigner, à pas réguliers (10 à 100 ms), la présence ou l'absence d'un type d'événement sonore [5]. Néanmoins, ces méthodes peuvent présenter des limites pour gérer des sources particulières, absentes de la base de données d'entraînement. La taille des réseaux ou la complexité des calculs peuvent aussi être un facteur limitant.

C'est pourquoi nous souhaitons développer une méthode alternative [7], qui soit capable de détecter les instants de début et de fin des événements sonores. Cette méthode correspond à l'utilisation de critères statistiques, destinés à détecter des ruptures au sein d'un signal temporel. Ces critères ne sont pas appliqués au signal acoustique brut, mais à un ensemble de composantes, qui mettent en avant les variations importantes.

De nombreux travaux, appliqués au traitement de la parole, utilisent les MFCCs [8, 9] pour détecter les changements d'interlocuteur. Néanmoins, ceux-ci sont sensibles à la superposition de sources et à la présence de bruit [9], deux phénomènes souvent retrouvés en acoustique environnementale [3]. Dans cet article, une nouvelle méthode de détection, basée sur les niveaux d'énergie en bandes de tiers d'octave, est présentée.

Le principe de la méthode de détection d'événements sonores proposée est d'abord présenté dans la Section 2. Les critères statistiques utilisés pour localiser/positionner les ruptures sont ensuite détaillés dans la Section 3. Les résultats de détection obtenus sur les signaux de notre base test sont enfin présentés et les intérêts de la technique proposée sont discutés.

2 Méthode de détection

On considère un signal $p(t)$, de durée T, représentant l'évolution temporelle de la pression acoustique. Ce signal est découpé en segments réguliers, de durée Δt (Fig. 1). Sur chaque segment, on calcule les niveaux d'énergie sur les N bandes de tiers d'octave, afin d'obtenir la matrice $X(t)$ (Eq. 1) :

$$X(t) = \left(x_1(t) \dots x_N(t)\right)^T \tag{1}$$

où $x_i(t), i \in [1:N]$ est l'évolution du niveau d'énergie sur la bande de tiers d'octave d'indice i.

On définit alors une fenêtre d'analyse, de durée T_{ev} (Fig. 1), qui est un ordre de grandeur de la durée minimale des événements sonores à détecter.

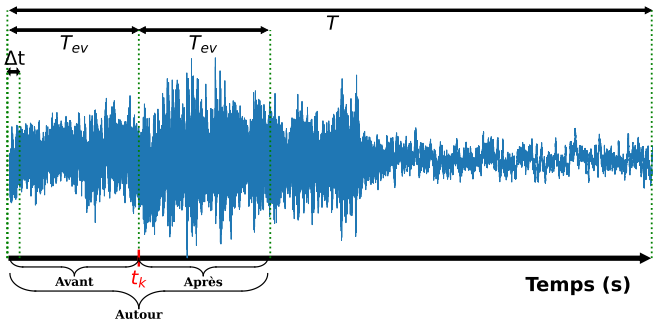

FIGURE 1 : Schéma explicatif des définitions des durées Δt, T_{ev} et T, ainsi que les intervalles "Avant", "Après" et "Autour", définis par rapport à t_k.

La moyenne μ et la matrice de covariance Σ des niveaux en bandes de tiers d'octave sont estimées pour chaque intervalle. Pour un ensemble Ω contenant n éléments, on a respectivement $\mu_\Omega = \frac{1}{n}\sum_{x_j\in\Omega} x_j$ et $\Sigma_\Omega = \frac{1}{n-1}\sum_{x_j\in\Omega}(x_j-\mu_\Omega)(x_j-\mu_\Omega)^T$

Nous obtenons alors 6 variables, qui sont traitées par les critères statistiques introduits dans la section suivante. Chaque critère en déduit un score, qui traduit la similarité entre les intervalles "Avant" et "Après" l'instant t_k.

Algorithme 1 : Algorithme de calcul de l'évolution du critère S

Entrées : $X(t)$: Évolution des niveaux d'énergie par bande de tiers d'octave.
T_{ev} : Durée des intervalles "Avant" et "Après".
Sorties : $S(t)$: Évolution temporelle du critère S.
pour $t_k = T_{ev}, T_{ev}+\Delta t, ..., T-T_{ev}$ **faire**
- Estimer la distribution des descripteurs "Avant" $t_k : \rightarrow (\mu_{k,av}; \Sigma_{k,av})$
- Estimer la distribution des descripteurs "Après" $t_k : \rightarrow (\mu_{k,ap}; \Sigma_{k,ap})$
- Estimer la distribution des descripteurs "Autour" de $t_k : \rightarrow (\mu_{k,au}; \Sigma_{k,au})$
- Calculer $S(t_k)$ à l'aide des paramètres $(\mu_{k,av}; \Sigma_{k,av}; \mu_{k,ap}; \Sigma_{k,ap}; \mu_{k,au}; \Sigma_{k,au})$

fin

L'algorithme 1 présente la méthode de calcul de l'évolution temporelle d'un critère, notée $S(t)$. Celle-ci est calculée avec un pas temporel de Δt, petit devant T_{ev}. De ce fait, une unique rupture réelle se traduit par une haute probabilité de rupture sur plusieurs échantillons successifs de $S(t)$. Pour éviter que l'algorithme détecte plusieurs fois une seule rupture, celles-ci sont positionnées à chaque maximum local de $S(t)$ [2, 8, 9]. Par construction, nos critères sont négatifs ou nuls dans une situation de continuité du signal. L'algorithme place donc une rupture à chaque maximum local positif d'un critère statistique, sans dépendre d'un seuil supplémentaire.

3 Présentation des critères

Une fois les estimations des distributions réalisées dans l'algorithme 1, il existe plusieurs manières de déterminer la présence ou non d'une rupture. Les études peuvent être réalisées en environnement calme (milieu rural) ou en environnement bruité, avec de multiples sources sonores qui se recouvrent (milieu urbain). Ainsi, nous aurons besoin de critères robustes, capables de détecter des événements qui émergent peu. Pour cela, on propose 4 manières différentes de traiter les distributions estimées, qui utilisent différentes propriétés pour évaluer les possibilités de rupture. Les particularités de chaque critère pour la tâche de détection sont par la suite discutées.

3.1 Bayesian Information Criterion (BIC)

Le BIC (2) [2, 6] s'appuie sur les calculs de log-vraisemblance associés aux matrices de covariance de chacun des ensembles "Avant t_k" ($\Sigma_{k,av}$), "Après t_k" ($\Sigma_{k,ap}$) et "Autour de t_k" ($\Sigma_{k,au}$).

$$\begin{aligned}\mathrm{BIC}(t_k) =& \frac{T_{ev}}{\Delta t}\left(\ln|\Sigma_{k,au}| - \ln|\Sigma_{k,av}| - \ln|\Sigma_{k,ap}|\right)\\ &- \ln\left(\frac{2T_{ev}}{\Delta t}\right)\end{aligned} \quad (2)$$

3.2 Distance de Kullback-Leibler (KL)

La distance de Kullback-Leibler (3) [2, 9] traduit une distance entre les sous-ensembles "Avant t_k" et "Après t_k". Les moyennes $(\mu_{k,av}; \mu_{k,ap})$ et matrices de covariance $(\Sigma_{k,av}; \Sigma_{k,ap})$ de ces deux ensembles sont utilisées, mais on ne s'occupe pas du grand ensemble "Autour" $(\mu_{k,au}; \Sigma_{k,au})$.

$$\begin{aligned}\mathrm{KL}(t_k) =& \frac{1}{2}\left(\left(\mu_{k,av}-\mu_{k,ap}\right)^T\left(\Sigma_{k,av}^{-1}+\Sigma_{k,ap}^{-1}\right)\left(\mu_{k,av}-\mu_{k,ap}\right)\right)\\ &+\frac{1}{2}\mathrm{tr}\left(\Sigma_{k,av}^{-1}.\Sigma_{k,ap}+\Sigma_{k,ap}^{-1}.\Sigma_{k,av}-2I\right)\end{aligned} \quad (3)$$

3.3 Statistique du T² (T²)

Le critère T² (4) [9] met en avant les potentielles différences de moyennes entre les deux sous-ensembles "Avant t_k" et "Après t_k" $(\mu_{k,av}; \mu_{k,ap})$ mais aussi les effets de variance sur l'ensemble "Autour de t_k" $(\Sigma_{k,au})$.

$$\mathrm{T^2}(t_k) = \frac{T_{ev}}{2\Delta t}\left(\mu_{k,av}-\mu_{k,ap}\right)^T.\Sigma_k^{-1}.\left(\mu_{k,av}-\mu_{k,ap}\right) \quad (4)$$

3.4 Distance euclidienne (d)

Le critère de la distance euclidienne (5) correspond simplement à la distance, composante par composante, entre les moyennes des sous-ensembles "Avant t_k" et "Après t_k" $(\mu_{k,av}; \mu_{k,ap})$.

$$\mathrm{d}(t_k) = \sqrt{\left(\mu_{k,av}-\mu_{k,ap}\right)^T.\left(\mu_{k,av}-\mu_{k,ap}\right)} \quad (5)$$

4 Évaluation de la méthode proposée

La méthode de détection proposée et les critères statistiques présentés sont évalués sur une base de signaux acoustiques. Cette base est composée de 12 enregistrements de 30 minutes en conditions réelles et 3 signaux synthétisés, pour un total d'environ 6 heures et 30 minutes de signal. Ces 12 enregistrements proviennent de 7 environnements différents, en milieu urbain ou rural. Les signaux sonores sont échantillonnés à $F_e = 25600$ Hz. Les bandes de tiers d'octave sont ainsi calculées de 20 Hz à 11.225 kHz, ce qui correspond à une large gamme des fréquences audibles par l'oreille humaine.

Les ruptures relatives à chaque critère de détection sont comparées aux délimitations de référence, sélectionnées par l'utilisateur. Deux métriques communes pour les travaux de détection d'événements sont appliquées pour évaluer les performances [5]. Il s'agit de la précision p et du recall r donnés par les équations suivantes :

$$\text{Précision} : p = \frac{TP}{TP+FP} \quad ; \quad \text{Recall} : r = \frac{TP}{TP+FN}$$

avec TP le nombre de vrais positifs, FP le nombre de faux positifs et FN le nombre de faux négatifs.

La précision évalue la quantité de fausses alarmes, le recall décrit si les ruptures de référence sont détectées.

5 Présentation des résultats

Nous travaillons d'abord sur un extrait de signal synthétisé, construit à partir d'un environnement calme et d'événements réels non perturbés. Sur l'extrait présenté sur la figure 2, deux passages de train (autour de $t = 10$ s et $t = 40$ s) ainsi que deux passages de voiture (avant $t = 30$ s et après $t = 50$ s) ont été rajoutés. Ces événements sont délimités par les bornes rouges sur le log-spectrogramme de la figure 2.

Pour appliquer nos critères, on choisit une fenêtre d'analyse de durée $T_{ev} = 2$ s, qui correspond à la durée de l'événement de référence le plus court. Les niveaux d'énergie sont calculés en bandes de tiers d'octave sur des segments de durée $\Delta t = 125$ ms, avec un recouvrement de 50%. Chaque critère est appliqué sur l'évolution des niveaux d'énergie de chaque bande de tiers d'octave. Cependant, la figure 2 illustre l'évolution du niveau sonore équivalent sur 125 ms, noté $\text{Leq}_{,\,125ms}$, calculé sur toutes les bandes de tiers d'octave considérées de 1 à N.

Sur cet exemple, chaque critère fournit une détection satisfaisante, hormis le dernier passage de voiture qui n'est pas détecté par le BIC. La délimitation exacte des événements varie selon les critères. Le BIC et la distance de Kullback-Leibler arrivent, par l'utilisation des matrices de covariance $\Sigma_{k,av}$ et $\Sigma_{k,ap}$, à détecter plus tôt l'arrivée d'une source en mouvement.

Néanmoins, le constat majeur sur ce premier exemple concerne la présence de nombreuses ruptures qui ne sont pas liées aux événements de référence. En effet, le traitement appliqué détecte tous les événements sonores présents, et cela peut concerner des sources peu émergentes. Ici, l'environnement comprend plusieurs sources émettant en basses fréquences, qui sont bien détectées par les critères. Hors, ces sources contribuent peu au bruit global, et leur détection peut alors constituer une information superflue. C'est pourquoi on cherche généralement dans le domaine de l'acoustique environnementale à caractériser plutôt une source sonore particulière ou bien les sources les plus émergentes.

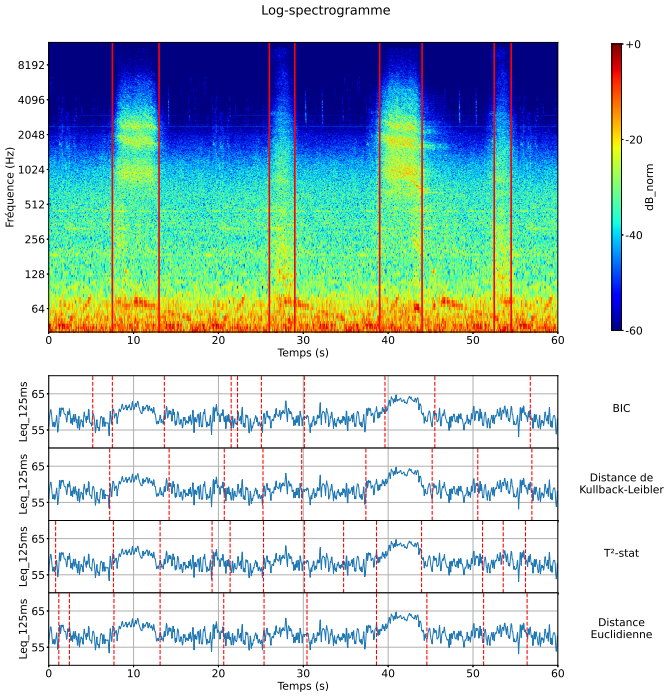

FIGURE 2 : Log-spectrogramme et ruptures détectées par les critères statistiques pour un signal synthétisé en environnement calme, avec des passages ferroviaires et routiers.

6 Application aux sources ferroviaires

En reprenant l'exemple étudié (cf. Fig. 2) et les mêmes paramètres Δt et T_{ev}, on souhaite détecter uniquement les passages ferroviaires. La connaissance de la source et de l'environnement permet d'anticiper les bandes de tiers d'octave caractérisant les passages ferroviaires. La figure 3 est l'application des critères sur la bande fréquentielle [1 kHz : 3.15 kHz], caractéristique de la source ferroviaire [4].

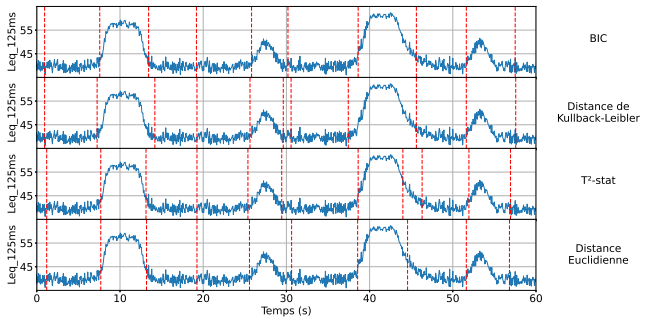

FIGURE 3 : Evolution des $\text{Leq}_{,\,125ms}$ calculés sur la bande [1 kHz : 3.15 kHz], et ruptures détectées par les critères sur cet intervalle de fréquence.

Sur la figure 3, la restriction en fréquences permet d'éliminer les perturbations en basses fréquences en conservant les passages de train. Cependant, les passages de voiture émergent aussi à ces fréquences, et sont toujours détectés.

En ajustant la durée T_{ev}, on peut séparer des sources sonores émettant dans les mêmes fréquences. Mais dans cet exemple, les durées des passages de train et de voiture sont équivalentes : il n'existe pas de durée T_{ev} qui élimine les passages de voiture en conservant les passages de train.

Dans ce cas, s'il est nécessaire de ne conserver que les passages du train, l'utilisation d'un seuil complémentaire est envisageable. Ce seuil, appliqué sur les critères, permet ici de distinguer les passages ferroviaires, mais il nécessite une analyse a posteriori.

La table 1 renseigne les métriques pour chaque critère appliqué à la base de données complète, qui comprennent 100 passages ferroviaires dans des contextes variés. 3 bandes de fréquences différentes sont utilisées, afin de représenter une adaptation progressive aux fréquences caractéristiques de la source ferroviaire.

TABLE 1 : Détection de trains avec les $\mathrm{L_{eq},\ }_{125ms}$ et $T_{ev} = 5$s.

Fréquences (Hz)	[20 :11.225k]	[400 :4k]	[1k :3.15k]
Precision BIC	0.15	0.20	0.18
Precision KL	0.23	0.18	0.18
Precision T²	0.15	0.16	0.17
Precision Eucl.	0.19	0.21	0.19
Recall BIC	0.79	**0.89**	**0.89**
Recall KL	0.36	**0.82**	**0.86**
Recall T²	0.93	0.94	0.95
Recall Eucl.	0.94	0.96	0.93

À l'instar de la figure 3, la précision plutôt constante montre que la restriction en fréquences ne permet pas d'extraire uniquement la source ferroviaire. Si les phénomènes en basses fréquences sont retirés, d'autres événements, émettant aux mêmes fréquences que les passages ferroviaires, peuvent être conservés voire introduits.

Concernant l'impact de la bande de fréquences choisie sur le recall, on remarque deux comportements différents selon les critères. La statistique du T² et la distance euclidienne, qui se concentrent sur l'évolution des moyennes, favorisent les sources à large bande fréquentielle. C'est pourquoi le recall est déjà élevé sans restriction en fréquences. Au contraire, le BIC et la distance de Kullback-Leibler, qui utilisent les matrices de covariance des sous-ensembles "Avant" et "Après", sont sensibles aux variations courtes ou sur des bandes de fréquence plus étroites. Sans adaptation en fréquences, le recall de ces critères est plus faible pour une source à large bande fréquentielle. Une fois les fréquences adaptées selon la source, le recall s'améliore nettement pour atteindre des valeurs intéressantes.

Les résultats avec des niveaux équivalents calculés sur 20 ms répètent globalement la même tendance, avec de meilleurs résultats. Dans les configurations les plus favorables, on a une précision proche de 0.33 pour un recall entre 0.9 et 0.97. Mais cette amélioration a un coût en termes de temps de calcul, proportionnel au nombre d'itérations nécessaires pour parcourir le signal entier.

7 Conclusion

La caractérisation d'une source sonore présente dans l'environnement passe par la détection de sa contribution. Dans cet article, une nouvelle méthode de détection d'évènements sonores est proposée. Elle repose sur l'analyse de différents critères statistiques, estimés à partir de l'évolution temporelle des niveaux d'énergie en bandes de tiers d'octave. Les ruptures obtenues correspondent à la détection de non-stationnarité sur les évolutions des niveaux d'énergie. Les résultats des tests soulignent la capacité de différents critères utilisés à détecter les évènements sonores au sein d'un signal acoustique. Or, en acoustique environnementale, les sources sont multiples et peuvent se recouvrir. L'apport d'informations supplémentaires améliore la détection d'une source particulière au sein d'un environnement complexe. Les performances des critères utilisés sont évaluées pour la tâche de détection d'une centaine d'événements sonores ferroviaires. La détection de cette source est favorisée par l'utilisation des critères de T² et de la distance euclidienne. Ces deux critères sont bien adaptés à des évènements large bande. Cependant, la distance de Kullback-Leibler et le BIC doivent être guidés, grâce à l'information des fréquences caractéristiques du bruit ferroviaire. En effet, l'adaptation de la bande fréquentielle d'analyse autour des fréquences ferroviaires permet d'améliorer les performances.

Malgré tout, pour isoler complètement les événements d'une seule source sonore, un traitement complémentaire doit être mis en œuvre. Nous avons vu que la caractérisation en fréquences permettait de corriger les situations de faible recall. Concernant la précision, l'application d'un seuil sur les critères peut permettre de se concentrer sur la source souhaitée et d'ignorer les perturbations mineures.

Références

[1] *Environmental noise guidelines for the European region*. World Health Organization. Regional Office for Europe, 2018.

[2] S. CHEN, P. GOPALAKRISHNAN *et al.* : Speaker, environment and channel change detection and clustering via the bayesian information criterion. *In Proc. DARPA broadcast news transcription and understanding workshop*, volume 8, pages 127–132. Citeseer, 1998.

[3] S. CHU, S. NARAYANAN et C.-C.J. KUO : Environmental sound recognition with time–frequency audio features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17:1142 – 1158, septembre 2009.

[4] Abbes KACEM : *Auralisation des transports ferroviaires en milieu urbain*. Thèse de doctorat, Université Grenoble Alpes (ComUE), 2019.

[5] A. MESAROS, T. HEITTOLA, T. VIRTANEN et M.D. PLUMBLEY : Sound event detection : A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83, septembre 2021.

[6] G. SCHWARZ : Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[7] C. TRUONG, L. OUDRE et N. VAYATIS : Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[8] J. ŽĎÁNSKỲ : Detection of acoustic change-points in audio streams and signal segmentation. *Radioengineering*, 2005.

[9] B. ZHOU et J.H.L. HANSEN : Efficient audio stream segmentation via the combined t²-statistic and bayesian information criterion. *IEEE Transactions on Speech and Audio Processing*, 13(4):467–474, 2005.